DOCUMENT RESUME

ED 410 275 TM 027 077

AUTHOR Crehan, Kevin D.
TITLE An Investigation of the Validity of Locally Developed Performance Measures in a School Assessment Program.
PUB DATE Mar 97
NOTE 14p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; Educational Assessment; *Educational Change; Elementary Education; Elementary School Curriculum; Elementary School Students; Language Arts; *Performance Based Assessment; Reading Tests; *School Districts; Standardized Tests; Standards; *Test Use; *Validity
IDENTIFIERS Authentic Assessment; Comprehensive Tests of Basic Skills

ABSTRACT

A large school district conducting a revision of its curriculum-based assessment program for grades one through six introduced performance assessments in response to a demand for more authentic assessment associated with national concerns for standards-based educational reform. This study reports validity evidence for the locally developed reading and language arts performance assessments, administered as part of the Curriculum- Based Performance Assessment (CBAP). The Comprehensive Tests of Basic Skills (CTBS) were used as the norm-referenced measure in this school system. Assessment results from third and fourth grade for over 6,000 students were used in the analysis. Correlations were determined for performance assessment and CTBS scores. Most notable were the low correlations among the performance assessment scores and CTBS and CBAP multiple choice scales. Inspection of the correlation matrices for both grades does not provide clear validity evidence for convergent or discriminant patterns. Results of this study may lead to questioning the value of performance assessments in a school district assessment program. It may be that the cost of preparing, administering, and scoring these assessments outweighs their benefits. (Contains 3 tables and 13 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Running head: An Investigation of Validity

ED 410 275

AN INVESTIGATION OF THE VALIDITY OF LOCALLY DEVELOPED PERFORMANCE MEASURES IN A SCHOOL ASSESSMENT PROGRAM

Kevin D. Crehan

University of Nevada, Las Vegas

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Kevin Crehan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Il, March, 1997.

TM027077

BEST COPY AVAILABLE

An Investigation of the Validity of Locally Developed Performance Measures in a School Assessment Program

A large school district conducted a thorough revision of its curriculum-based assessment program in grades one through six. The revision was necessary to bring the assessment program in line with revised curricula and to supplement the revised multiple-choice tests with performance assessments. The introduction of performance assessments was in response to the growing interest in more "authentic" assessment associated with national concerns for standards-based educational reform. Following intensive preparation and field-testing activities, the revised multiple-choice tests and performance measures were administered for the first time as the district's Curriculum-Based Performance Assessment (CBAP).

The introduction of the new performance measures into the assessment program raises a number of questions. An in-depth analysis of the questions involved in the implementation of performance assessments is presented by Linn, Baker, and Dunbar (1991). They counsel that the issue of evaluating the quality of alternative assessments has not been sufficiently considered and suggest eight criteria for judging the adequacy of performance-based assessments: consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost and efficiency. Additionally, Messick (1995) discusses six aspects of construct validation for performance assessments: content, substantive, structural, generalizability, external, and consequential.

While all these criteria are important and need to be addressed in some manner by the school district, the present focus is on the external aspect of construct validity. More specifically, how do the new performance assessments relate to the revised curriculum-based multiple-choice tests and to the state-mandated Comprehensive Tests of Basic Skills, Fourth Edition (CTBS/4) (1989) achievement test battery?

It appears that most of the attention to the psychometric quality of performance assessments has focused on reliability (e.g., Lane, Stone, Ankenmann, & Liu, 1992; Linn & Burton, 1994; Linn, 1993; Shavelson, Baxter, and Pine, 1992) with evidence of validity limited to judgments of content adequacy, quality, and complexity. While there has been work in language arts assessment relating so called "direct" and "indirect" measure of language artss, little evidence on the relationship between widely used norm-referenced multiple-choice measures of achievement and performance assessments used in large-scale school assessment programs was found. Burger and Burger (1994) used a multimethod-multitrait approach (Campbell & Fiske, 1959) to assess the validity of the Michigan Education Assessment Program (MEAP) and a locally developed language arts assessment using the CTBS/4 as the criterion assessment of achievement. Moderate validity evidence to support the external aspect (Messick, 1995) was observed for both performance measures. Correlations for the language arts assessment were attenuated due to low inter-rater agreement.

This study reports validity evidence similar to the Burger and Burger (1994) study.

The CTBS/4 is used as the norm-referenced achievement measure along with locally developed reading and language arts performance assessments. An investigation of correlations among the measures will allow evaluation of convergent and discriminant validity evidence among the new performance assessments and other measures of achievement which utilize multiple-choice item formats.

The appropriateness of using a norm-referenced multiple-choice measure of achievement as the validity criterion is arguable. However, these tests are valued by educators (Center for Research on Evaluation, Standards, and Student Testing, 1990) and are useful in the prediction of achievement and as general indicators of achievement in specific content areas (Worthen & Spandel, 1991).

METHODS

Subjects

Assessment results from third and fourth grade for over 6000 students were utilized in the analysis. These results include both the third and fourth grade spring administrations of the locally developed CBAP multiple-choice and performance measures described above and third grade fall results for the CTBS/4 (1989).

Instruments

The instruments used in this investigation are described in the following. The CBAP multiple-choice tests are locally developed curriculum-based measures of reading

comprehension, language arts, and mathematics (3rd and 4th grade forms). Locally developed CBAP performance assessments are designed to yield measures of response to reading, language arts, and mathematics (3rd and 4th grade forms). Two similar sets of two reading/language arts were developed for third and fourth grade. One set was provided for instructional practice a week or two before the actual spring assessment administration. Correspondingly, two sets of two similar mathematics assessments were developed for each grade level. Again, one set of these assessments was provided for instructional practice prior to the assessment administration. Each student was administered one of the two performance assessments in reading/language arts and math. The CTBS/4 (1989) was administered in October of the third grade.

Procedure

The results of these testings were match-merged to allow correlations among the measures to be determined. The CTBS/4, CBAP multiple-choice, and CBAP performance assessments each yield scores for reading, language (or language arts), and math. Correlations were determined among these scores for interpretation in the manner of a multitrait-multimethod matrix separately for third and fourth grade.

RESULTS

Table 1 reports the summary descriptive statistics and reliabilities for the measures used in this study. No reliability estimates for the performance assessments are given since each student responded to only one of the assessments.

Table 2 presents the multitrait-multimethod correlation matrix for the third grade measures. The monotrait correlations are underlined and the monomethod correlations are enclosed in triangles. The remaining correlations are between different traits and different methods. Most notable in this matrix are the low correlations among the performance assessment scores and both the CTBS/4 and CBAP multiple-choice scales. This observation will be discussed later. Convergent evidence of validity would be confirmed by the monotrait correlations being the highest and discriminant evidence would be substantiated by the remaining different trait and different method correlations being the lowest. Since the monotrait correlations are not uniformly higher than the monomethod correlations, evidence of convergence is absent. Additionally, since there is no clear pattern of the different trait and different method correlations being the lowest, discriminant evidence of validity is also lacking. Table 3 presents the multitrait-multimethod correlations for the fourth grade CBAP results with the third grade CTBS/4 scores. Again, correlations among the performance assessments and the CTBS/4 and CBAP multiple-choice measures are low. As with the third grade results, inspection of the matrix fails to confirm clear validity evidence for either convergent or discriminant patterns.

DISCUSSION

The increasing use of performance measures in educational assessment programs suggests the need for more empirical evidence of the relationship of these newer measures to those measures with which educators have greater familiarity. While Burger and Burger (1994) found some validity evidence for the performance assessments used in their study, the present study fails to replicate their results using different performance assessments. Partial

explanation of the absence of evidence for the external aspect of construct validity may be due to the score coding method used by the school district. Although students took one of two different forms of the performance assessments in reading/language arts and math, the score coding did not differentiate between the two assessments. If different forms of the performance assessments were not similar in difficulty, the correlations would be attenuated. Another potential explanation for the low performance assessment correlations may be attributable to modest inter-rater agreement and generalizability. Inter-rater agreement and G-study coefficients for the reading assessments used in this study were observed earlier (Crehan, Hudson, & Costa, 1994). Inter-rater agreement among five trained raters was in the 20% range for the same score and in the 70% range for ratings within one point. G-study coefficients were in the .60 range.

While performance assessment score coding and low inter-rater agreement are potential problems in this study, the limitation placed on curriculum content sampling with performance assessments may be a more serious concern. Even given perfect inter-rater agreement, the error associated with task sampling limits the reliability of these measures. In preparing the performance assessments, attempts were made to tap a number of curriculum elements in each of the areas of reading, language arts, and math. However, the sampling was clearly far narrower than for the multiple-choice tests, especially in the area of mathematics.

The results of this study may lead one to question the value of performance

assessments in a school district assessment program. It may be that the cost of preparing, administering, and scoring these assessments outweighs their benefits. Perhaps a better use of this assessment method would be at the classroom level. Classroom teachers could administer and score the assessments and include the results in the student's portfolio to aid in tracking progress. Following the assessment, teachers could also give students feedback on their performance and encourage revision toward a standard of mastery.

It has been suggested that using the performance assessment more as a regular feature of instruction than as a test may enhance student learning. Shepard et al. (1996) observed the effects on reading and mathematics achievement of including performance assessments in classroom instruction with third graders. Although the evidence from this one-year study is equivocal, it is suggested that study of the effects of incorporating performance assessments into the instructional program be continued. Well constructed performance assessments are designed to stimulate the thinking processes related to important learning outcomes. Additionally, the outcome of the exercise is the translation of the thinking into a synthetic product, e.g., written or oral expression. The value of this type of instructional activity seems obvious and compelling. Perhaps a longer trial period will yield more encouraging results.

REFERENCES

Burger, S. E., & Burger, D. L. (1994). Determining the validity of performance-based assessment. Educational Measurement: Issues and Practice, 13(1), 9-15.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Center for Research on Evaluation, Standards,, and Student Testing. (1990). Monitoring the impact of testing and evaluation innovations project: State activity and interest concerning performance-based assessment. ERIC document ED327570.

Comprehensive Test of Basic Skills. (1989). CTB MacMillan/McGraw-Hill, Monterey, CA.

Crehan, K.D., Hudson, R., & Costa, J.S. (1994). Introducing locally developed performance measures into a school assessment program. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Lane, S., Stone, C.A., Ankenmann, R.D., & Liu, M. (1992). Empirical evidence for the reliability and validity of performance assessments. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20, 15-21.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. Educational Measurement: Issues and Practice, 13(1), 5-8, & 15.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. Educational Evaluation and Policy Analysis, 15, 1-16.

Messick, S. (1995). Standards of validity and the validity standards in performance assessment. Educational Measurement: Issues and Practice, 14(4), 5-8.

Shavelson, R. J., Baxter, G. P., & Pine J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 21, 22-27.

Shepard, L.A., Flexer, R.J. Hiebert, E.H., Marion, S.F., Mayfield, V., & Weston, T.J. (1996). Effects of introducing classroom performance assessments on student learning. Educational Measurement: Issues and Practice, 15(3), 7-18.

Worthen, B. R., & Spandel, V. (1991, February). Putting the standardized test debate in perspective. Educational Leadership, 65-69.

Table 1

Number of Items, Means, Standard Deviations, and KR20's for all Instruments. Sample sizes exceed 6000 students.

Variable	Grade	#Items	Mean	SD	KR20
CTBS/4 Read	Gd 3	40	22.9	7.29	.85
CTBS/4 Lang	Gd 3	40	22.4	7.13	.85
CTBS/4 Math	Gd 3	40	19.8	6.73	.80
CBAP Read	Gd 3	55	45.1	7.89	.90
CBAP Lang	Gd 3	45	37.2	5.82	.83
CBAP Math	Gd 3	60	45.9	9.57	.91
PA Read	Gd 3	1	1.8	.75	--
PA Lang	Gd 3	1	2.3	.61	--
PA Math	Gd 3	1	3.6	1.27	--
CBAP Read	Gd 4	53	40.8	8.56	.89
CBAP Lang	Gd 4	60	46.0	9.17	.91
CBAP Math	Gd 4	60	42.8	10.02	.90
PA Read	Gd 4	1	2.02	.80	--
PA Lang	Gd 4	1	2.26	.54	--
PA Math	Gd 4	1	2.66	1.17	--

Table 2

Multitrait-Multimethod Correlation Matrix for Grade Three Monotrait-Multimethod Correlations are Underlined and Multitrait-Monomethod Correlations are Outlined with a Triangle. Sample Size Exceeds 6000 Students.

		(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) CTBS/4	Read	.71	.59	.62	.59	.57	.20	.33	.05
(2) CTBS/4	Lang		.64	.59	.62	.58	.22	.38	.04
(3) CTBS/4	Math			.51	.54	.61	.19	.31	.05
(4) CBAP	Read				.69	.68	.26	.37	.09
(5) CBAP	Lang					.68	.24	.37	.08
(6) CBAP	Math						.24	.36	.12
(7) PA	Read							.33	.05
(8) PA	Lang								.08
(9) PA	Math								

Table 3

Multitrait-Multimethod Correlation Matrix for Grade Four

Monotrait-Multimethod Correlations are Underlined and Multitrait-Monomethod Correlations are Outlined with Triangles. CTBS/4 Results are from Third Grade. Sample Size Exceeds 6000 Students.

		(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) CTBS/4	Read	.71	.59	.65	.65	.58	.25	.31	.21
(2) CTBS/4	Lang		.64	.63	.68	.60	.26	.35	.33
(3) CTBS/4	Math			.55	.56	.63	.24	.28	.25
(4) CBAP	Read				.77	.71	.28	.35	.27
(5) CBAP	Lang					.72	.29	.38	.27
(6) CBAP	Math						.27	.33	.31
(7) PA	Read							.41	.16
(8) PA	Lang								.15
(9) PA	Math								